



CARDIFF  
UNIVERSITY



PRIFYSGOL  
CAERDYDD

Cardiff University  
Computer Science and Informatics

CM3202 – Individual Project (40 Credits)

Initial Plan

Author:  
Laura Edwards

Supervisor:  
Christopher Jones

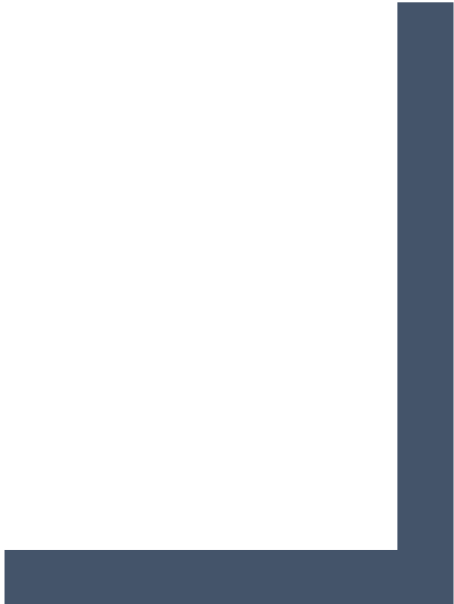


Table of Contents

1. PROJECT TITLE..... 2

2. PROJECT DESCRIPTION ..... 2

3. PROJECT AIMS & OBJECTIVES ..... 3

3.1. MAIN AIMS AND OBJECTIVES ..... 3

3.2. DESIRABLES ..... 5

4. WORK PLAN ..... 5

4.1. DELIVERABLES ..... 5

4.2. SUPERVISOR MEETINGS ..... 5

4.3. RESEARCH ..... 6

4.4. MILESTONES ..... 6

REFERENCES..... 9

## 1. Project Title

“Species distribution modelling with machine learning”

## 2. Project Description

Recent years have seen a growing interest in species distribution modelling resulting from the rapid development occurring in this area. The increasing demand for information has led to innovative and effective techniques and data suitable for addressing this information need e.g. statistical learning methods.

The aim of this project is to utilize machine learning methods in order to predict the geographical distribution of specific wildlife species in the UK. Machine learning stems from Artificial intelligence to elicit information from given datasets. Making Machine Learning algorithms a promising method when predicting and modelling species distribution. Species distribution combines two kinds of data, data on species occurrence and environmental data.[1]

I will be experimenting with similar methods to those used in papers by Jeawak et al, in which social media data were used in addition to conventional environmental data sources. However, I'll be looking at other sources of data than those used in these paper by Jeawak et al. The “Using Flickr for characterizing the environment: an exploratory analysis” paper used both environmental data and Flickr, not the NBN data I intend to use for species distribution.[2] The Jeawak et al paper entitled “Mapping Wildlife Species Distribution with Social Media: Augmenting Text Classification” did use NBN data but it only used Flickr for prediction without using any other environmental data.[3]

In order to complete this project I will acquire a number of datasets which can be extracted using available API's and stored appropriately. I intend to divide the UK into grid cells and will use data from the citizen science portal such as the National Biodiversity Network [4] on particular species as the ground truth of whether the particular species is observed giving the presence and absence in each cell. Data to characterise and differentiate between locations will be obtained from various sources

that record environmental features such as climate, land cover, soil type. These datasets come in many formats such as raster vector and will need to be prepared into a format which the model will be able to interpret and use.

When implementing machine learning classifiers I will utilise the NBN data by dividing the ground truth data into training, tuning and testing (this data is not used for training). The test data is used to evaluate the performance of the methods / algorithms, while the training data is being used to train the machine learning classifiers. I intend to develop similar classifiers to those mentioned in Jian Zhangs paper entitled “A review of machine Learning Based Species’ Distribution Modelling”. Examples of the machine learning applied include Random Forests, Support Vector Machine (SVM).[1] I will be implementing these algorithms using my chosen programming language of python and with the aid of python libraries such as ‘sklearn’.

[5]

This project does not only entail me experimenting with different classifiers but also diverse combinations of environmental features and individually to find the least and most effective characteristics when predicting the occurrence of specific wildlife distribution.

### 3. Project Aims & Objectives

With just a 12 week time limit given to complete this project I have split my aims and objects into main and desirables. Main aims and objectives are the minimum to which must be achieved to complete the project. Desirables are those that I would like to achieve if time allows.

#### 3.1. Main Aims and Objectives

This projects aim is to use different machine learning classifiers algorithms and identify the one that performs best at predicting species distribution. For this project I will choose 4 different types of classifiers to develop and evaluate performance. Furthermore, a comparison of combinations of environmental features will be performed. Appropriate feature selection will enable the developed classifier to

significantly increase its performance. The classifier algorithm will be implemented using existing python libraries.

- Research machine learning classifiers
  - Research similar projects that use classifiers and which are appropriate to apply such as Naïve Bayes, Support Vector Machine(SVM) and Random Forests.
  - Research suitable ways to visualise the results.
  - Research python libraries available to help with implementing algorithms such as 'sklearn'.
- Research environmental features
  - Study characteristics and gain understanding of important environmental features that affect wildlife patterns that I should include. Such as climate, Land cover etc.
- Research species
  - I intend to focus on a number of particular species or a group.
  - This requires research as there needs to be enough data available for that species.
- Collect and store datasets
  - Collect ground truth data from National Biodiversity Network API
  - Collect environmental features data from various sources.
  - Prepare datasets in a format for which the model will be able to interpret the data.
- Implement and use machine learning to successfully predict species distribution
  - Develop a comprehensive comparison between 4 different machine learning algorithms with a suitable visualisation of the results.
  - Compare different combinations of environmental features to predict the occurrence of particular species.
  - Find the most effective classifier with the optimal combination of features for predicting species distribution..
- Final Report
  - Assess the milestones I set out to achieve and analyse how I progressed through the stages.

- Collate notes from major stages made in diary log and expand in detail how models developed.
- Summarise how my objectives came together to complete the project.

### 3.2. Desirables

- Deep learning approaches [6]
  - Research and implement one of the various models available, such as the Bert model from Hugging Face [7]
- Compare different size grid cells
  - Collect further data on selected environmental features for 2 other grid cell sizes and compare results.

## 4. Work Plan

### 4.1. Deliverables

Deliverables for the project include the initial plan (22<sup>nd</sup> February), due to a 2 week extension following extenuating circumstances. For the final deadline on the 10<sup>th</sup> May I must submit a final report which includes research, sample code and evaluation of my work. By the 10<sup>th</sup> May all my code developed must be refined and ready to submit along with supportive documents, graphs and tables.

- Initial Report
- Final Report
- All source code
- Supporting Documents
- Graphs and tables

### 4.2. Supervisor Meetings

I have discussed with my supervisor Chris and arranged a weekly meeting on a Thursday. This is going to ensure my work rate remains consistent, along with structured help, guidance and support on the project.

### 4.3. Research

I have given myself plenty of time for research as a lot of the topics are unfamiliar to myself, thus will need to build this knowledge. This project is heavily reliant on collecting appropriate data so I plan to do this first.

I am going to gauge below the time scale for all the coding and research I am intending to complete. This is my initial plan which may not be accurate if there are unforeseen circumstances that will cause me to deviate from this. I fortunately do not have to encounter any exams this semester which would have affected the amount of work I would have been able to complete that week. With only one other module this semester I intend to keep Easter free for if there is any work I need to catch up on for that module. I have allocated a week for desirables which are tasks that I have mentioned prior that I wish to complete if time allows. I intend to keep a diary every week that keeps a log of what I have done, my approach, if I achieved my goal for that week, what went well and how it could be improved, this will help me towards writing my final report.

### 4.4. Milestones

- Week 1-2 (*1<sup>st</sup> Feb – 12<sup>th</sup> Feb*)
  - Initial meeting with project supervisor to discuss initial plan.
  - Research projects on species distribution modelling
  - Research Machine learning classifiers
  - Draft for initial report
- Week 3 (*15<sup>th</sup> Feb – 19<sup>th</sup> Feb*)
  - Send first draft to supervisor for feedback
  - Write final draft of initial plan
  - Research on Machine learning classifiers
  - Research environmental characteristics that affect wildlife
  - Weekly Supervisor meeting
  - **Deliverables: Initial Plan**
- Week 4 (*22<sup>nd</sup> Feb – 26<sup>th</sup> Feb*)
  - Download and install relevant software and libraries.
  - Collect datasets. Both NBN and environmental characteristics

- Research implementation of the 4 selected classification algorithms possibly using 'sklearn'.
  - Weekly Supervisor Meeting
  - Write Diary Log
- Week 5 (*1<sup>st</sup> March – 5<sup>th</sup> March*)
  - Develop 2 of the 4 models and evaluate their effectiveness.
  - Weekly Supervisor Meeting
  - Write Diary Log
- Week 6 (*8<sup>th</sup> March – 12<sup>th</sup> March*)
  - Develop the other 2 models and evaluate their effectiveness.
  - Weekly Supervisor Meeting
  - Write Diary Log
- Week 7 (*15<sup>th</sup> March – 19<sup>th</sup> March*)
  - Apply different combinations of features on the 4 models that have been developed and see how the classifiers performance is affected.
  - Weekly Supervisor Meeting
  - Write Diary Log
- Week 8 (*22<sup>nd</sup> March – 26<sup>th</sup> March*)
  - Evaluate which classifier with combination of features gives the best performance.
  - Produce graphs to show the different classifiers and features
  - Refine code and debug
  - Weekly Supervisor Meeting
  - Write Diary Log

***Easter break***

- Week 9 (*19<sup>th</sup> April – 23<sup>rd</sup> April*)
  - Finalize on work and complete any tasks that may not have been finished
  - Allocated time for desirables
  - Weekly Supervisor Meeting
  - Write Diary Log
- Week 10 – 11 (*26<sup>th</sup> April – 30<sup>th</sup> April + 3<sup>rd</sup> May – 7<sup>th</sup> May*)
  - Initial draft of final report
  - Refine code where needed



- Add appropriate comments to code for readability
  - Weekly Supervisor Meeting
- Week 12 (*10<sup>th</sup> May – 14<sup>th</sup> May*)
  - Complete final draft of final report
  - Hand-in final report
  - **Deliverables: Final report, source code, graphs**

## References

- [1] J. Zhang and S. Li (2017). "*A Review of Machine Learning Based Species' Distribution Modelling*," International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), Wuhan, China, 2017, pp. 199-206
- [2] Jeawak, S. S., Jones, C. B., Schockaert, S., 2017. *Using Flickr for characterizing the environment: an exploratory analysis*. In: *13th International Conference on Spatial Information Theory*, COSIT 2017. Vol. 86 of Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, pp. 21:1–21:13.
- [3] Jeawak, S.S., C.B. Jones, S. Schockaert (2018) '*Mapping wildlife species distribution with social media: Augmenting text classification with species names*'. GIScience 2018. Leibniz International Proceedings in Informatics.
- [4] NBN Atlas. Available at: <https://species.nbnatlas.org>
- [5] Scikit-learn. Available at: <https://scikit-learn.org/stable/>
- [6] Botella C., Joly A., Bonnet P., Monestiez P., Munoz F. (2018) *A Deep Learning Approach to Species Distribution Modelling*. In: Joly A., Vrochidis S., Karatzas K., Karppinen A., Bonnet P. (eds) *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*. Multimedia Systems and Applications. Springer, Cham.
- [7] Hugging Face. Available at: <https://huggingface.co/models>
- [8] Science Advances. Available at: <https://advances.sciencemag.org/content/5/1/eaat4858>
- [9] Scott Jarvie and Jens-Christian Svenning. *Using species distribution modelling to determine opportunities for trophic rewilding under future scenarios of climate change*

[10] Syed Amir Manzoor, Geoffrey Griffiths and Martin Lukac. *Species distribution transferability and model grain size – finer may not always be better.*